Erschienen in: Schwitalla, Johannes/Wegstein, Werner (Hrsg.): Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003. - Tübingen: Niemeyer, 2005. S. 57-70.

Rainer Perkuhn / Cyril Belica / Doris al-Wadi / Meike Lauer / Kathrin Steyer / Christian Weiß

Korpustechnologie am Institut für Deutsche Sprache

1. Einleitung

Das Institut für Deutsche Sprache (im folgenden kurz *IDS* genannt) hat die Aufgabe, den Gebrauch der deutschen Sprache der Gegenwart und der jüngeren Vergangenheit zu dokumentieren und zu erforschen. Aufgrund der Möglichkeiten, die die elektronische Datenverarbeitung heutzutage bietet, drängt es sich geradezu auf, reale Texte digital aufzubereiten, zu speichern und darüber hinaus Anwendungssoftware zu entwickeln, mit deren Hilfe die Dokumentation und Erforschung eine neue Qualität erhält.

Dieser Bericht gibt einen kurzen Überblick über die Tätigkeiten im Bereich Korpustechnologie am IDS, soweit diese in den Zuständigkeitsbereich der Arbeitsgruppe (AG) für Korpustechnologie fallen. Dadurch bedingt konzentriert er sich auf den Bereich der geschriebenen bzw. verschriftlichten deutschen Sprache der jüngeren Gegenwart und setzt den Schwerpunkt auf die Tätigkeiten seit 1992. Der Vollständigkeit halber seien das Recherchesystem COSMAS II, die Arbeiten im Bereich historischer Korpora und im Bereich der Korpora gesprochener Sprache erwähnt, die allerdings in eigenständigen Projekten angesiedelt sind.

Arbeitsgruppe für Korpustechnologie

Die Arbeitsgruppe für Korpustechnologie befasst sich mit Weiter- und Neuentwicklungen im IDS und weltweit genutzter Korpustechnologien. Ihre Aufgaben sind

- der Ausbau und die Pflege der Korpora geschriebener Gegenwartssprache sowohl durch laufende Akquisition neuer Texte als auch durch Erarbeitung von Quellenbibliografien, Entwicklung von Algorithmen zur Qualitätsprüfung und -sicherung der Korpora, Entwicklung von Verfahren zur Verbesserung der automatischen Dokumentation der Korpora – als Daueraufgabe gemäß dem Auftrag des Instituts –, sowie

- die Erforschung und Entwicklung von korpusorientierten Erschließungsmethoden und darauf basierenden anwenderfreundlichen Analysetechniken, z.B. statistische Methoden zur Erfassung und Auswertung zeit- und textsortenrelevanter Parameter, Kookkurrenzanalyse, Erfassung kontextspezifischer semantischer Phänomene, Zeitreihenanalysen zur Gewinnung von Neulexemen für die Neologismenforschung, Weiterentwicklung der Lemmatisierungsverfahren, Gewinnung von frequenzattribuierten Lemmaregistern aus den Korpora.

2. Textkorpora des IDS — empirische Basis für die linguistische Forschung —

Die Korpora geschriebener Gegenwartssprache des IDS

bilden mit knapp zwei Milliarden Wörtern die weltweit größte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit.
sind gebührenfrei zugänglich über COSMAS, das IDS-eigene Recherche- und Analyse-System, das speziell auf linguistische Bedürfnisse abgestimmt ist.
enthalten belletristische, wissenschaftliche und populärwissenschaftliche Texte, eine große Zahl von Zeitungstexten sowie eine breite Palette weiterer Textarten und werden kontinuierlich weiterentwickelt.
werden im Hinblick auf große Variabilität und Menge akquiriert und erlauben in der Nutzungsphase über COSMAS die Komposition virtueller Korpora, die repräsentativ oder auf spezielle Aufgabenstellungen zugeschnitten sein können.

2.1 Korpusakquisition – Vom Originaltext zum Korpustext

Textbeschaffung

Die Korpora geschriebener Gegenwartsprache sollen den tatsächlichen Gebrauch der deutschen Sprache dokumentieren und diese Dokumentation stetig, am besten täglich, fortschreiben. Das heißt, als mögliche Quellen kommen künstliche Texte nicht und Webseiten nur bedingt in Frage, da sie nur einen sehr speziellen Ausschnitt der Sprache darstellen. Die Aufgabe ist demnach die Beschaffung von geeigneten, möglichst elektronischen Vorlagen, die ein authentisches Dokument des Gebrauchs der deutschen Sprache darstellen. Germanisten und andere Forscher benötigen dieses Material, um Sprache wissenschaftlich empirisch zu erforschen. Insofern sollte es eigentlich selbstverständlich, ja sogar eine Ehre sein, dass Autoren ihre Vorlagen zur Verfügung stellen. Die Rechteinhaber werden aber häufig von der Angst abgeschreckt, mit der Freigabe ihrer Texte für die Korpora lasse sich deren widerrechtliche Vervielfältigung nicht mehr kontrollieren. Darüber hinaus sind die meisten Vorlagen aus technischen Gründen nicht besonders gut geeignet. Die Textproduzenten sind selten willens oder in der Lage, ihre Quellen in einem Format aufzubereiten, das eine leichte Überführung in das Korpusformat ermöglicht. Andererseits verfügt das IDS nicht über die Kapazitäten, die ansonsten erforderliche aufwändige Aufbereitung der Quellen selber zu übernehmen. Aufgrund dieser Rahmenbedingungen gestalten sich die rechtlichen und finanziellen Verhandlungen meistens sehr schwierig.

Urheberrechte

Durch juristische Vereinbarungen mit Verlagen, Zeitungsredaktionen und Autoren war und ist das IDS in der Lage, urheberrechtlich abgesichertes Textmaterial derart zu beschaffen, dass alle Korpora IDS-intern und Teile dieser Korpora weltweit öffentlich genutzt werden können, und zwar ausschließlich zu wissenschaftlichen, nichtkommerziellen Zwecken. Die Textkorpora des IDS sind zudem nur über das COSMAS-System recherchierbar; kein Nutzer hat Zugriff auf vollständige Korpustexte, sondern nur auf begrenzte Kontexte zu Suchanfragen.

Deutsches Referenzkorpus

Das vom Land Baden-Württemberg finanzierte Kooperationsprojekt *Deutsches Referenzkorpus* (DEREKO) begann im Mai 1999 und endete im März 2002. Kooperationspartner des IDS waren das *Institut für Maschinelle Sprachverarbeitung* (IMS) der Universität Stuttgart und das *Seminar für Sprachwissenschaft* (SfS) der Universität Tübingen. Ziel war, die deutsche Gegenwartssprache (von 1956 bis Ende 2001) möglichst breit und der Sprachwirklichkeit angemessen zu repräsentieren, mit modernen korpuslinguistischen Verfahren aufzubereiten und der Wissenschaft zu Verfügung zu stellen. Das Projekt war in die Gesamtvorhaben der Arbeitsgruppe für Korpustechnologie eingebunden, schloss unmittelbar an die Korpusakquisitionsarbeiten des IDS in den vergangenen Jahren an und nutzte die Ressourcen innerhalb des IDS.

Das Projekt hat viel zur Klärung gerade der urheberrechtlichen Unsicherheit beigetragen. Der Anteil der Korpora, die öffentlich zugänglich sind, stieg durch dieses Projekt beträchtlich.

Es hat sich gezeigt, dass Korpusakquisition, -aufbereitung und -dokumentation als eine Daueraufgabe aufgefasst werden muss, damit der Wandel der deutschen Sprache in Wortschatz und auch Grammatik kontinuierlich dokumentiert und erforscht werden kann. Das Projekt hat ferner deutlich gemacht, dass die genannten kommunikativen, juristischen, textlinguistischen, technologischen und bibliografischen Arbeiten beim Aufbau eines Referenzkorpus einen hohen personellen Aufwand erfordern.

Konvertierung

Die Quelltexte, die in die Korpussammlung aufgenommen werden sollen, liegen normalerweise in fremden Formaten vor, die auf die Bedürfnisse des Publikationswesens zugeschnitten sind und die je nach Präferenzen des Autors und des Verlags stark variieren können. Um Teil der IDS-Korpora zu werden, müssen sie in ein einheitliches, durch das IDS-Textmodell beschriebenes Format überführt werden. Das bedeutet, dass große Mengen sehr heterogener Daten in mehreren Arbeitsschritten analysiert und aufwärts konvertiert werden müssen. Zur maschinellen Unterstützung werden dazu verschiedene Parser, Konvertierer und Filter z.T. selbst entwickelt und eingesetzt. Probleme bereiten z.B. Inkonsistenzen bei Zeichensatzkodierung und Worttrennung sowie ‚Datenmüll' wie unmotivierte Formatangaben, Tabellen oder Textdopplungen. Aber auch sinnvolle Formatangaben variieren je nach Quelle und müssen für eine optimale Datenüberführung eingehend analysiert und vereinheitlicht werden. Die Aufwärtskonvertierung der Quellen in das IDS-Format hat einen stark iterativen Charakter und ist wegen des damit verbundenen hohen Korrektur- bzw. Wartungsbedarfs sehr kosten- und zeitaufwändig.

2.2 Umfang

Das IDS begann Mitte der Sechzigerjahre mit dem Aufbau elektronischer Textkorpora. Der Umfang der Korpora hat sich seit 1992 von ca. 28 Millionen auf über 1,9 Milliarden Textwörter im Jahre 2002 erhöht (das entspricht etwa 4.800.000 Buchseiten).

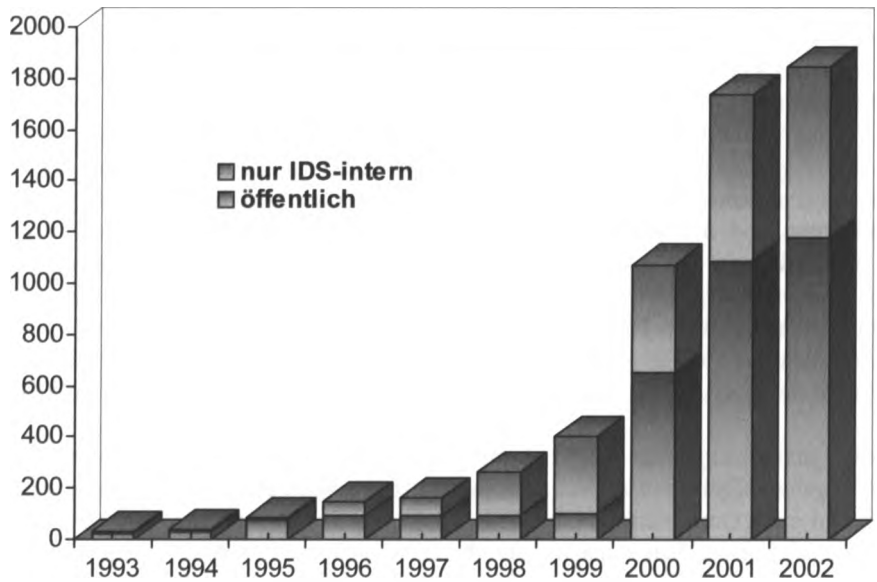


Abb. 1: Größe der COSMAS-I-Korpora (in Millionen Textwörtern)

2.3 Morphosyntaktisch annotierte Korpora

Ein Teil der IDS-Korpora wurde Mitte der Neunzigerjahre automatisch morphosyntaktisch annotiert. Diese Arbeiten wurden im Jahr 1996 bis zur Fertigstellung des COSMAS-II-Systems eingestellt. Die bereits annotierten Korpora (ca. 26 Millionen Textwörter) wurden 1998 ersatzweise über das COSMAS-I-System zur Nutzung freigegeben.

2.4 Das IDS-Textmodell

Für die effiziente automatische Auswertung großer elektronischer Textsammlungen ist es unerlässlich, die Texte in einem einheitlichen Format zur Verfügung zu haben. Dieses Format wird in dem TEI[1]-/ CES[2]-konformen IDS-Textmodell definiert. Charakteristisch für das IDS-Textmodell sind die angestrebte originalgetreue Abbildung der textuellen Inhalte und Strukturen der Quelltexte und die Dokumentation sämtlicher bisher in den Korpora vorkommender Textarten in einheitlichen Strukturen.

Zentrale Komponenten des IDS-Textmodells sind die Korpusstruktur, die Korpustext-Bibliografie und die Primärtextbehandlung.

---

[1] Text Encoding Initiative: http://www.teic.org/Guidelines/index.htm

[2] Corpus Encoding Standard: http://www.cs.vassar.edu/CES

Korpusstruktur
Um virtuelle Korpuskompositionen, sinnvolle Quellenlisten bei der Ergebnispräsentation u.a.m. zu ermöglichen, werden die Quelltexte nach festgelegten Kriterien gegliedert und in eine hierarchische Struktur eingebunden, die folgende drei Ebenen umfasst:

↳ Korpusebene (Korpusidentifikator, z.B. LES)
↳ Dokumentebene (Dokumentidentifikator, z.B. LES/ESS)
↳ Textebene (Textidentifikator, z.B. LES/ESS.20022)

Das IDS-Textmodell definiert Text als eine relativ selbstständige, inhaltlich kohärente Folge natürlichsprachlicher Äußerungen, die natürlichen Kommunikationssituationen entstammen. Sie bildet den Korpustext, die »kleinste« Einheit eines Korpus.

Jedes Korpus besteht aus einem oder mehreren Dokumenten; jedes Dokument setzt sich wiederum aus einem oder mehreren Korpustexten zusammen. In einem Dokument können mehrere Texte nach bestimmten Gesichtspunkten zusammengefasst sein, z.B. nach Quellen, chronologischer Abfolge, Themenbereichen und/oder Textarten. Ein Text beinhaltet je nach Korpusstruktur z.B. einen oder mehrere Zeitungsartikel oder eine als Ganzes aufgenommene Zeitung/Zeitschrift, einen Auszug aus einem selbstständigen Werk oder ein selbstständiges Werk als Ganzes.

Beispiel: Das Korpus Siegfried Lenz: Werkausgabe in Einzelbänden [20 Bde.]. – Hamburg: Hoffmann und Campe Verlag, 1996-1999

Texte	Dokument	Beschreibung	Bd
1	LES/HIL.00000	Es waren Habichte in der Luft. Roman	1
...	...	...	...
1	LES/ALE.00000	Die Auflehnung. Roman	12
77	LES/ERZ.13001[-16022]	[Erzählungen]	13-16
3	LES/SCH.17001[-17003]	[Schauspiele]	17
4	LES/HOR.18001[-18004]	[Hörspiele]	18
98	LES/ESS.19001[-20032]	[Essays]	19+20

Korpustext-Bibliografie
Die IDS-Korpustexte sind von jeher mit Quellennachweisen versehen, die bei der Anzeige gefundener Belege mit angezeigt werden. Allerdings waren sie in den früheren Korpora unstrukturiert. So wurde in den Neunzigerjahren ein Korpustext-Bibliografiemodell als eine zentrale Komponente des IDS-Textmodells entwickelt, das korpusübergreifende automatische Zugriffe auf die nunmehr einheitlich strukturierten umfangreichen Quellendaten mit folgenden Zielen erlaubt:

automatische virtuelle Korpuskomposition nach Autoren, Textarten, Entstehungszeiten, Sachgebieten usw.; vorkommende Textarten sind z.B.:

- Abhandlung
- Aphorismus
- Aufsatz
- Autobiografie
- Bericht
- Biografie
- Brief
- Denkschrift
- Erlass
- Erzählung
- Essay
- Flugblatt
- Fußnote
- Forschungsbericht
- Gebet
- Gebrauchsanweisung
- Gedicht
- Handzettel
- Hörspiel
- Interview
- Klappentext
- Leitartikel
- Märchen
- Nachruf
- Nachwort
- Parteiprogramm
- Petition
- Presseerklärung
- Produktbeschreibung
- Protokoll
- Rede
- Rezension
- Roman
- Schauspiel
- Tagebuch
- Vorspann
- Werbung

automatische nutzerorientierte Generierung von auswählbaren Arten von Quellennachweisen (ausführlich normgerecht, verkürzt oder übergeordnet),
Informationsgewinnung statistischer Natur unter vielfältigen Aspekten, z.B. chronologische Sortierung der Rechercheergebnisse, ermöglicht durch die Bereitstellung des Entstehungsdatums (s. Grafik).

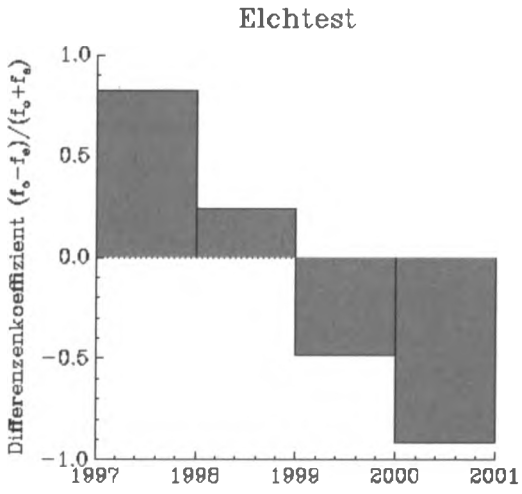


Abb.2: Gebrauchshäufigkeitsveränderung des Wortes Elchtest

Primärtextbehandlung
Der Primärtext des IDS-Textmodells setzt sich aus den so originalgetreu wie möglich abgebildeten Originaltext- und minimalen notwendigen Zusatzinformationen zusammen.

Originaltextinformationen = Grundtext + Vorspann + Überschrift(en)
+ Grußformel(n) + Bildtext(e) + Zusatz-(Rand-)Text(e)
+ Übersicht(en) + Tabelle(n) + Fußnote(n)
+ Originalseiteneinteilung + ...
Zusatzinformationen = Satzende(n) + Absatzgrenze(n)

Das Markieren dieser Erscheinungen und ggf. weiterer (z.B. Verfasser, Interviewpartner, typografische Hervorhebungen) erlaubt die Inbezugsetzung von Textinhalten zu Textstrukturen, Recherchen mit Satzabständen, die Bestimmung der Belegstellenkontexte, die Bereitstellung konkreter Originalseitenangaben in den Quellennachweisen u.a.m.

3. Methoden der Korpusanalyse und –erschließung. Erforschung und Entwicklung von korpusanalytischen Verfahren

Korpora werden zwar häufig Volltextdatenbanken genannt, sie sind aber nur bis zu einem gewissen Grad mit Datenbanken vergleichbar. Während die Inhalte einer Datenbank weitestgehend formal strukturiert, z.T. auch typisiert sind, ist der wesentliche Bestandteil eines Korpus natürlichsprachlicher Text. Dieser kann zwar mit zusätzlichen, strukturierten Informationen angereichert, etwa bibliografisch annotiert sein, er kann auch über Layout und/oder Gliederung eine gewisse Struktur haben – diese Informationen sind aber normalerweise nicht diejenigen, die ein Linguist erfragen möchte. Von Interesse sind gerade Erkenntnisse über Phänomene der Sprache, die in den Texten leider nicht strukturiert vorliegen. Deshalb sind neue Verfahren zu entwickeln, die den Nutzern helfen, die Korpora zu erschließen und zu analysieren.

Für Belegrecherchen (im weitesten Sinne) ist eine spezielle, diesen Nutzern gerechte Anfragesprache erforderlich. Die Konzepte dafür – inklusive der Lemmatisierung und der Möglichkeiten, die morphosyntaktische Annotationen eröffnen – wurden in der Arbeitsgruppe entwickelt und in COSMAS I umgesetzt. Das Nachfolgesystem COSMAS II, für dessen Programmierung und Betrieb die Arbeitsstelle Zentrale Datenverarbeitung zuständig ist, hat die Konzepte übernommen und um eine grafische Darstellung erweitert. Die Arbeitsgruppe beschäftigt sich auch weiterhin mit der methodisch-konzeptuellen Weiterentwicklung des Korpusrecherchetools.

Ein herkömmliches Datenbanksystem stellt idealerweise genau die Daten zur Verfügung, die in die Datenbank eingegeben wurden, nicht weniger, aber normalerweise auch nicht mehr. Aus der Datenbankforschung heraus hat sich aber inzwischen eine eigene Disziplin entwickelt, die sich gerade zum Ziel gesetzt hat, neue nicht-triviale Information in den riesigen Datensammlungen zu entdecken: Data Mining bzw. Knowledge Discovery in Databases. Ihr Ziel ist es, neue Zusammenhänge und Strukturen zu erkennen, die einem menschlichen Betrachter nur schwer zugänglich sind. Datenbanken mit den entsprechenden Informationen vorausgesetzt, können diese Verfahren z.B. aufdecken, dass sich Krankmeldungen zu bestimmten Zeiten (etwa bei einer Fußballweltmeisterschaft), aus bestimmten Wohnvierteln oder in bestimmten Abteilungen häufen. Wohlgemerkt können

diese Verfahren nur zahlenmäßige Auffälligkeiten feststellen. Mit einer anschaulichen Darstellung, einer Visualisierung, kann die Interpretation dieser Auffälligkeiten unterstützt werden, indem der zeitliche Verlauf in einem Diagramm oder Krankmeldungen als Punkte auf einen Stadtplan oder ein Organigramm projiziert werden. Dadurch treten die Phänomene nicht nur quantitativ, sondern auch für einen unbedarften Benutzer leicht wahrnehmbar zu Tage.

Zwar verfolgt die Arbeitsgruppe ähnliche Ziele, doch erfordern die besonderen Eigenschaften der Korpora die Erforschung und Entwicklung von neuen, korpusorientierten Erschließungsmethoden und darauf basierenden anwenderfreundlichen Analysetechniken, wie statistische Kookkurrenzanalysen, Verfahren zur Visualisierung und zur lexikologisch-lexikografischen Erschließung von Ergebnissen von Kookkurrenzanalysen, thematische Erschließung von Texten und Dokumentclustering, Zeitreihenanalysen für die Neologismenforschung und quantitative Analysen der deutschen Lexik.

3.1 COSMAS

Die Arbeitsgruppe für Korpustechnologie stellt der Öffentlichkeit die Methoden der Korpusanalyse und -erschließung integriert in einem komplexen Online-System, COSMAS, zur Verfügung. Von 1992 bis 2003 wurde COSMAS I angewendet und in der Zeit ständig von der Arbeitsgruppe weiterentwickelt. Es stand seit 1993 (über den WWW-Zugang seit 1997) auch auswärtigen Nutzern weltweit gebührenfrei zur Verfügung. Nach fast zwölf Jahren Produktionsbetrieb wurde das COSMAS-I-Projekt zum 7. März 2003 beendet und der Korpusrechercheservice wurde vom COSMAS-I-Nachfolgesystem COSMAS II übernommen. Das neue System fällt allerdings nicht mehr in den Zuständigkeitsbereich der Arbeitsgruppe, sondern wird von der Arbeitsstelle Zentrale Datenverarbeitung programmiert und betreut.

Entwicklung der COSMAS-Nutzung

Seit 1992 ist die IDS-interne und die externe Nutzung des COSMAS-Systems stetig gestiegen. Der Leistungszuwachs ohne Personalerweiterung wurde möglich durch systematische Verallgemeinerung und Ausrichtung der Funktionalität und der Benutzeroberfläche auf linguistische Bedürfnisse.

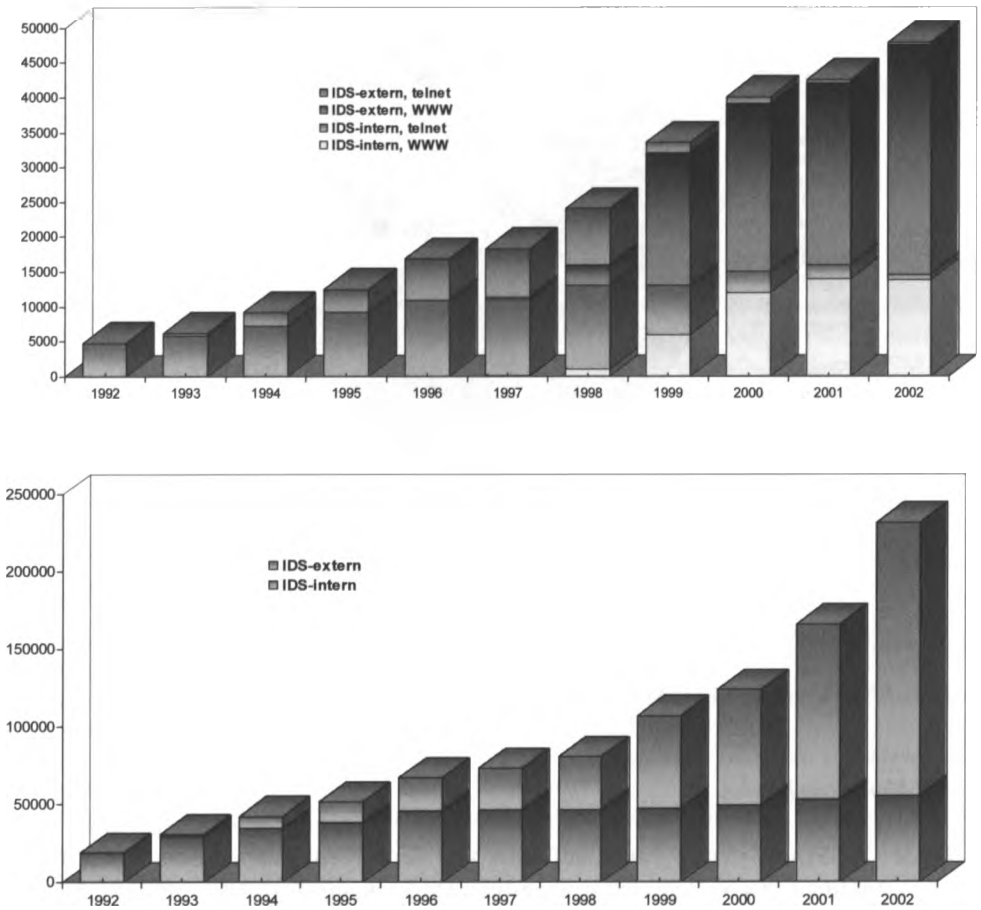


Abb. 3: Anzahl der COSMAS-I-Sitzungen

Seit Februar 2002 ist für den Zugang zu COSMAS eine Benutzerregistrierung erforderlich. In den ersten elf Monaten haben sich mehr als 4500 Benutzer für COSMAS I registrieren lassen, davon ca. 49% aus der Internet-Domäne .de (Deutschland). Die Anteile anderer Internet-Domänen sind aus der folgenden Grafik ersichtlich.

Wesentliche Leistungen von COSMAS, die auf von der AG entwickelten Konzepten und z.T. implementierten Modulen beruhen, sind

schnelle Recherchemöglichkeit in sehr großen Textkorpora,
ständige Verfügbarkeit,
korpusgerechte Lemmatisierung,
Kookkurrenzanalyse.

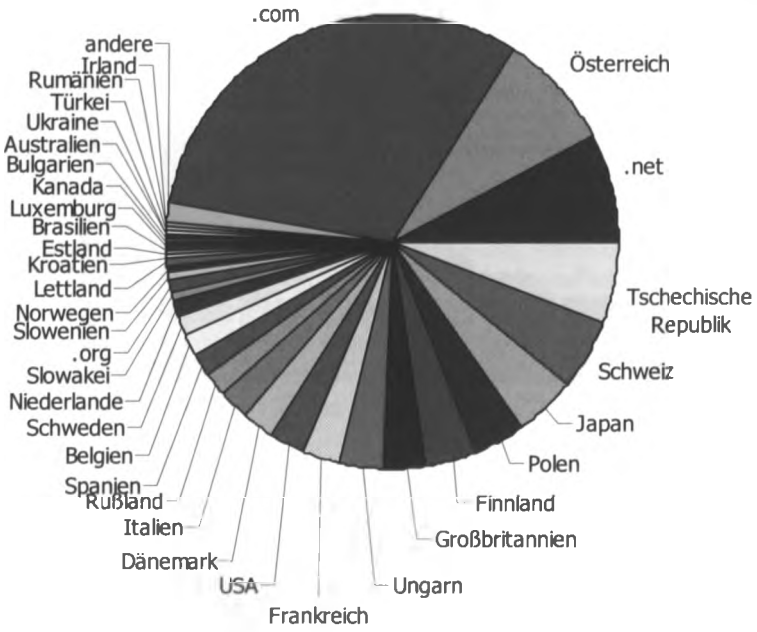


Abb. 5: Herkunft von COSMAS-I-Benutzern (nach Internet-Domänen ohne .de)

3.2 Korpusrecherche

Beide COSMAS-Systeme bauen auf einer von der Arbeitsgruppe konzipierten, reichhaltigen, auf linguistische Bedürfnisse abgestimmten Suchanfragesprache auf und ermöglichen

Recherche in vordefinierten oder frei zusammengestellten virtuellen Korpora,
Suche nach Wortformen und -segmenten, nach Satz- bzw. Sonderzeichen und Zahlen,
Suche mit Hilfe von Suchmustern (z.B. Grundformen) und Anzeige aller im ausgewählten Korpus belegten Wortformen eines Suchmusters, so dass relevante Formen selektiert werden können,
Einbeziehung morphosyntaktischer Merkmale und früherer Rechercheergebnisse in eine Suchanfrage,
Verknüpfung der verschiedenen Bestandteile über logische Operatoren sowie Abstandsoperatoren,
verschiedene Sortier- und Präsentationsmöglichkeiten, z.B. grafische Darstellung der chronologischen Sortierung,
Anzeige der Quellenangaben, der KWICs und der Volltexte,
verschiedene Exportmöglichkeiten.

Neben der Lemmatisierung sind es vor allem die morphosyntaktischen Merkmale, die es erlauben, spezifisch linguistische Anfragen zu formulieren.

3.3 Lemmatisierung

Die Lemmatisierung ermöglicht, dass Flexionsformen, Zusammensetzungen und/oder sonstigen Wortbildungsformen ihre Grundformen zugeordnet werden können. In diesem Zusammenhang sind Grundformen

unflektierte Simplizia verschiedener Wortarten,
unflektierte Ableitungen und Komposita,
Wortbildungsmorpheme.

Das Lemmatisierungsverfahren *Flexionsanalyse und Kompositazerlegung* zeichnet sich vor allem dadurch aus, dass es den besonderen korpustechnologischen Anforderungen gerecht wird. Seit 1994 wird eine Implementation des Verfahrens als ein Modul des COSMAS-Systems eingesetzt.

3.4 Kookkurrenzanalyse

Die Kookkurrenzanalyse ist eine korpusanalytische Methode zur Strukturierung von Belegmengen. Sie

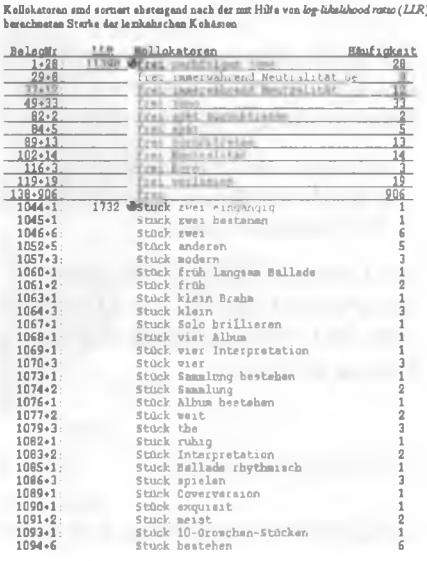
ermöglicht das Aufdecken von signifikanten Regelmäßigkeiten bei der Verwendung von Wortkombinationen in den Korpora,
wertet mit Hilfe mathematisch-statistischer Analyse- und Clusteringverfahren den definierbaren Kontext eines vorgegebenen Suchobjekts in beliebigen virtuellen Korpora aus,
liefert Hinweise auf systematisches gemeinsames Auftreten von Wörtern (Partnerwörter, Kollokatoren) und ein Maß für deren Affinität (Kohäsion),
fasst Belege, die ähnliches Kookkurrenzverhalten des Wortes dokumentieren, zu Gruppen/Clustern zusammen,
strukturiert diese Belegmengen ggf. hierarchisch,
bietet eine entsprechende synoptische Präsentation der Belege,
erfasst neben binären Wortrelationen auch usuelle phrasale Muster bis hin zu (idiomatischen) Mehrworteinheiten.

Die Arbeitsgruppe für Korpustechnologie stellt der Öffentlichkeit die Kookkurrenzanalyse seit 1995 integriert in COSMAS zur Verfügung. Die Kookkurrenzanalyse ist auf beliebige COSMAS-Suchobjekte anwendbar mit

optionaler Lemmatisierung,
variabler Kontextgröße,
ggf. automatischer Fokussierung auf den Kontext mit dem stärksten Kohäsionswert,
variabler Zuverlässigkeit (d.h. Signifikanz des ersten Kollokators),
variabler Granularität (d.h. Signifikanz der Kollokatoren, die für die Ermittlung von Mehrworteinheiten berücksichtigt werden),
variabler Zuordnung von Belegen bei Mehrworteinheiten.

Abb. 6: Kookkurrenzanalyse (in COSMAS I):

a) Anfrage



b) Clusterstruktur

c) KWICs

Beispiel einer Anwendung der Kookkurrenzanalyse

Am IDS findet die Kookkurrenzanalyse vor allem im Rahmen des Moduls ‚Usuelle Wortverbindungen' des Projekts *elexiko*, des IDS-Projekts zur korpusbasierten Erstellung eines elektronischen Informationssystems, seine Anwendung. Die empirische Methode

eröffnet einen empirischen Zugang zu Massendaten, indem sie durch Häufigkeitsbewertungen und Präferenzsetzungen hochfrequente Belegmengen ordnet und strukturiert,

dient darüber hinaus als heuristisches lexikografisches Arbeitsinstrument, indem es Evidenzen für aktuelle Sprachgebrauchsphänomene liefert, z.B. Informationen zu aktuellen Bedeutungen, typischen Verwendungsmustern, typischen grammatischen Mustern oder Sprachwandelerscheinungen,

ermöglicht eine Erfassung und Validierung usueller Wortverbindungen auf einer umfassenden empirischen Basis, um sie z.B. als Mehrworteinheiten (Kollokationen, Phraseologismen, Redewendungen, Sprichwörter, kommunikative Formeln, Funktionsverbgefüge usw.) der deutschen Gegenwartssprache lexikografisch aufbereiten zu können.

3.5 Kurzstudien

Zum Aufgabenspektrum der Arbeitsgruppe gehören ebenfalls zahlreiche korpusanalytische Machbarkeits- und Kurzstudien für Kollegen im IDS sowie für externe Wissenschaftler verschiedener Einrichtungen, z.B. Klinik für Psychosomatik und Psychotherapie, Justus-Liebig-Universität Giessen; Institut für Angewandte Sprachwissenschaft, Universität des Saarlandes; Linguistische Datenverarbeitung/Computerlinguistik an der Universität Trier; Lehrstuhl für Psychologie, Universität Mannheim; Universitätsklinik für Epileptologie in Bonn; Ernst Klett Verlag, Stuttgart; Sogang University, Seoul, Korea; Universität Zürich.

3.6 Aktuelle Teilprojekte

Zur Zeit starten folgende Teilprojekte der Arbeitsgruppe:

CCDB — die COSMAS-Kookkurrenzdatenbank
Für die Weiterentwicklung der Methoden der Kookkurrenzanalyse ist es von grundlegender Bedeutung, die zur Zeit noch weitestgehend unbekannten Eigenschaften von Kohäsionsrelationen zwischen Wörtern oder Wortgruppen der deutschen Sprache möglichst weit aufzudecken und zu systematisieren. In Zusammenarbeit mit dem Modul ‚Usuelle Wortverbindungen' des IDS-Projekts *elexiko* wird als empirische Basis dafür auf der Grundlage eines Korpus von ca. 1,6 Milliarden Textwörtern eine Kookkurrenzdatenbank zu den 10.000 lexikografisch komplexesten deutschen Wörtern erstellt. Neben der Erforschung der Eigenschaften von Kohäsionsrelationen für die Weiterentwicklung von Analysemethoden eignet sich die Datenbank natürlich auch als Nachschlagewerk bei der lexikografischen Arbeit. So kann man zum Beispiel - unter Berücksichtigung des zugrundeliegenden Korpus und der gewählten Analyseparameter - schnell und einfach auf Informationen zum Kookkurrenzverhalten einzelner Lexeme zugreifen. Für diese Zwecke steht die CCDB - COSMAS-Kookkurrenzdatenbank hausintern auch anderen Projekten zur Verfügung.

Lexikologische und lexikografische Erschließung der Kookkurrenzanalyse
Ziel des Teilprojekts ist es, Unterstützung für die lexikologische und lexikografische Erschließung der Kookkurrenzanalyse anzubieten, um die Vielfalt der Informationen sowohl einzelner als auch einer Menge von Kookkurrenzanalysen handhabbar zu machen. Der

Ansatz umfasst die Visualisierung der kohäsiven Struktur und Stärke der Kollokatoren, die Möglichkeit des Fokussierens einzelner Bereiche bzw. des Navigierens in einzelne Bereiche der visualisierten Struktur, verschiedene Möglichkeiten der lexikologischen und lexikografisch-redaktionellen Nachbearbeitung, sowie eine Schnittstelle zur CCDB.

Thematische Erschließung der Korpora
Ziel des Teilprojekts ist die thematische Erschließung der Korpora, um sowohl themenspezifische virtuelle Subkorpora zusammenstellen zu können als auch aufgrund der Analyse sachgebietsbezogener Häufigkeitsverteilungen z.B. Lesarten disambiguieren zu können. Zur semi-automatischen Gruppierung thematisch ähnlicher Wörter bzw. Korpustexte und zur Indexierung werden Techniken des *information retrieval* bzw. *document clustering* sowie lernbasierte Textklassifikatoren evaluiert und weiterentwickelt.

Danksagung
An dieser Stelle möchten wir allen danken, die dazu beigetragen haben, dass die IDS-Korpora der Gegenwartssprache die weltweit größte Sammlung dieser Art geworden sind. Eine ausführliche Dokumentation ist von Seiten der Arbeitsstelle für Öffentlichkeitsarbeit des IDS in Vorbereitung. Eine vorläufige Liste der Mitwirkenden finden Sie in unserer Webpräsentation.

Danken möchten wir auch allen Autoren, Verlagen und Zeitungen, die ihre Werke, Texte und Dokumente für die IDS-Korpora zur Verfügung gestellt haben und auch in Zukunft stellen werden. Gleichzeitig möchten wir alle bisher Abgeneigten und Unentschlossenen dazu ermuntern, sich an diesem Unterfangen zu beteiligen, ein möglichst breites und tiefes Abbild der deutschen Sprache der germanistischen Forschung zur Verfügung zu stellen. Falls Sie Fragen zum Urheberrecht und zur Vertragsgestaltung haben, informieren wir Sie gerne.